

## PRIMER NOTE

# Microsatellite markers developed from *Theobroma cacao* L. expressed sequence tags

JAMES W. BORRONE,\* J. STEVEN BROWN,\* DAVID N. KUHN,\* JUAN C. MOTAMAYOR† and RAYMOND J. SCHNELL\*

\*United States Department of Agriculture, †Masterfoods USA (Mars Incorporated), c/o United States Department of Agriculture, Agricultural Research Service, Subtropical Horticulture Research Station, 13601 Old Cutler Road, Miami, FL 33158, USA

## Abstract

*Theobroma cacao* L. expressed sequence tags (ESTs) were converted into useful genetic markers for fingerprinting individuals and genetic linkage mapping. Primers were designed to microsatellite-containing ESTs. Twenty-two *T. cacao* accessions, parents of various mapping populations segregating for disease resistance and crop yield characteristics, were tested. Twenty-seven informative loci were discovered with 26 primer pairs. The number of detected alleles ranged from two to 11 and averaged 4.4 per locus. All 27 markers could be mapped into at least one of the existing  $F_1$  or  $F_2$  populations segregating for agronomically important traits.

**Keywords:** expressed sequence tags, microsatellite, simple sequence repeats, *Theobroma cacao*

Received 10 July 2006; revision accepted 21 August 2006

*Theobroma cacao* L. is an important cash crop throughout the tropics (Souza & Dias 2001). The beans are the sole source of cocoa, the raw material for chocolate. Crop yields of *T. cacao* are limited by susceptibility to numerous pathogens. Conventional breeding programs for improved disease resistance have been hampered by the use of a narrow genetic base (Motamayor *et al.* 2002, 2003), mislabelling of accessions in international germplasm collections (Turnbull *et al.* 2004), and pollen contamination of controlled crosses (Cervantes-Martinez *et al.* 2006). Thus, traditional breeding has provided less than superior planting material for farmers (Dias 2001). Because of the small population sizes commonly produced in *T. cacao*, a haplotype-based method for quantitative trait loci (QTL) mapping in half-sibling  $F_1$  populations was recently proposed (Cervantes-Martinez & Brown 2004). The method requires that informative, codominant markers be present in the QTL-donor parent. To increase the number of microsatellites, publicly available *T. cacao* expressed sequence tags (ESTs) were investigated. Data-mining ESTs for microsatellites has proven effective for generating

novel markers in a number of plant species (Varshney *et al.* 2005).

*Theobroma cacao* ESTs (6557) were assembled into 2336 unigenes  $\geq 150$  bp in length using SEQUENCHER 4.0 (Gene Codes) and screened using the Cotton Microsatellite Database Simple Sequence Repeat Server (Clemson University Genomics Institute; [http://www.mainlab.clemson.edu/cmd/ssr\\_server/](http://www.mainlab.clemson.edu/cmd/ssr_server/)) with the default parameters for minimum repeat numbers. Polymerase chain reaction (PCR) primers were designed using PRIME in GCG (Wisconsin package version 10.2). Twenty-two accessions representing at least one founding parent of various  $F_1$  and  $F_2$  populations segregating for tolerance to vascular streak dieback, frosty pod, black pod or witches' broom diseases, as well as for important crop yield traits were assayed (Cervantes-Martinez *et al.* 2006).

DNA samples were extracted as described in Schnell *et al.* (2005). PCRs were conducted with PTC-225 thermal cyclers (MJ Research) containing 0.1  $\mu$ M forward primer, 0.1  $\mu$ M reverse primer, 200  $\mu$ M dNTPs, 10 ng of bovine serum albumin, 1  $\times$  PCR buffer with 1.5 mM MgCl<sub>2</sub>, 0.4 U of polymerase (either AmpliTaq, Applied Biosystems or NEB polymerase, New England Biolabs), and 10 ng of DNA in a total volume of 10  $\mu$ Ls. For primers with an

Correspondence: Raymond J. Schnell, Fax: 305-969-6410; E-mail: rschnell@saa.ars.usda.gov

annealing temperature ( $T_a$ ) of 60 °C, thermalcycler conditions were: 94 °C, 2 min; (94 °C, 20 s; 60 °C, 45 s; 72 °C, 1 min) × 5; (94 °C, 20 s; 56 °C, 45 s; 72 °C 1 min) × 35; 72 °C, 10 min. The first five cycles were omitted for primers with a  $T_a$  of 56 °C. Amplification success was determined by agarose gel electrophoresis: 2.5% agarose, 1 × TBE buffer, 120 V, ethidium bromide-stained, and visualized with UV light. Capillary electrophoresis was performed on the ABI PRISM 3100 Genetic Analyser with ROX-labelled GENESCAN-400HD (Applied Biosystems) as described in Schnell *et al.* (2005) except the injection time was reduced to 10 s if the signal strength of the amplified product saturated the detector. Forward primers were labelled with either 6-FAM or HEX. Alleles were scored using GENESCAN version 3.7 and GENOTYPER version 3.7 (Applied Biosystems). Descriptive statistics (Table 1) were generated with GDA version 1.1 (Lewis & Zaykin 2001). Tests for Hardy–Weinberg equilibrium (HWE) and linkage disequilibrium (LD) were conducted using GENEPOL version 3.4 (Raymond & Rousset 1995).

Twenty-seven polymorphic loci were identified from 26 primer pairs. One primer pair, SHRSTc047, amplified an additional locus designated SHRSTc047a. The number of alleles ranged from two to 11, averaging 4.4 alleles per locus (Table 1). Eleven loci departed significantly from HWE ( $P < 0.05$ ), and all showed significant LD ( $P > 0.05$ ) with at least two other loci. The observed departure from HWE and the LD noted is likely due to the nature of the accessions tested: cultivated material selected and used to develop populations segregating for resistance to various diseases as well as other agronomically important traits (pod number, bean size, etc.).

Fingerprinting of the accessions indicated 27 loci characterized could be placed in at least one mapping population (Table 2). The primers were tested on TSH516, an  $F_1$  individual used to create the only  $F_2$  population of *T. cacao* (Brown *et al.* 2005), and progeny of the  $F_2$  population and various  $F_1$  populations. Segregation at the expected ratios has been observed (data not shown). These EST-derived microsatellite markers are currently being mapped into these populations. Additional microsatellite-containing ESTs are currently being evaluated for their utility as genetic markers.

## References

- Brown JS, Schnell RJ, Motamayor JC *et al.* (2005) Resistance gene mapping for witches' broom disease in *Theobroma cacao* L. in an  $F_2$  population using SSR markers and candidate genes. *Journal of the American Society for Horticultural Science*, **130**, 366–373.
- Cervantes-Martinez C, Brown JS (2004) A haplotype-based method for QTL mapping of  $F_1$  populations in outbred plant species. *Crop Science*, **44**, 1572–1583.
- Cervantes-Martinez C, Phillips-Mora W, Brown JS *et al.* (2006) Combining ability for disease resistance, yield, and horticultural traits of cacao (*Theobroma cacao* L.) clones. *Journal of the American Society for Horticultural Science*, **131**, 231–241.

**Table 1** Primer sequences and characteristics of 27 *Theobroma cacao* microsatellite loci developed from ESTs

Locus	GenBank Accession†	Primer sequence (5'-3')	$T_a$ (°C)	Repeat(s)	n	Size range (base pairs)	No. of alleles	$H_E$	$H_O$	f	LD‡
SHRSTc045	CA794528	F: GAGCGAAAATGGCACAC R: GAGGTCATCCCTGAATCCAT	60	(GCA) <sub>7</sub> (ACA) <sub>6</sub> N <sub>3</sub> (CAA) <sub>2</sub> G <sub>2</sub> (AGA) <sub>4</sub>	22	222–228	3	0.635	0.727	0.149	60, 63, 67, 69
SHRSTc046	CA794695	F: GCGGCTACCTTCACTCT R: ATCAAAACAGACGCAA-AA	60	(TA) <sub>3</sub> (AT) <sub>8</sub> N <sub>2</sub> (CT) <sub>5</sub>	21	198–204	4	0.761	0.429*	0.443	—
SHRSTc047	CA794728	F: AAAGGAAATCAGAGAGAG R: ACTTGAGGAATTGGAACT	56	(AGA) <sub>7</sub>	22	142, 154	2	0.333	0.318	0.045	50, 70
SHRSTc047a	unknown			unknown	22	225–234	3	0.604	0.682	-0.133	50, 52, 59, 60, 67, 68
SHRSTc048	CA794970	F: AAATCCCCCTGGTTCTACTCC R: GATTCGAGTGTAAAAGTGTG	60	(CAT) <sub>4</sub> N <sub>27</sub> (CAT) <sub>3</sub> N <sub>2</sub> (TCA) <sub>2</sub> N <sub>3</sub> (TCA) <sub>7</sub>	21	278–284	3	0.570	0.714	-0.261	50, 53, 56, 61, 64, 70
SHRSTc049	CA797725	F: ATCCGAGCAAACCTCCCTCTC R: TTCTCTTCCCACCAAGTCCC	60	(CTCCCT) <sub>4</sub>	22	263–284	8	0.642	0.455*	0.296	—
SHRSTc050	CA797995	F: GAAAGGGGG2ATGAG R: GCAGATGGAAACAGGGAT	60	(TTO) <sub>2</sub> N <sub>4</sub> (GAA) <sub>7</sub> (GCA) <sub>3</sub>	22	217–241	5	0.754	0.773	-0.026	47, 47a, 48, 61, 66, 67, 68, 70
SHRSTc051	CA798018	F: CTGGTTTTCGCTCCCTTGCT R: ATTGCTGGTTCTCCATCT	60	(TAA) <sub>2</sub> (AT) <sub>3</sub> (AG) <sub>2</sub> N <sub>2</sub> (TA) <sub>11</sub>	22	177–186	7	0.652	0.727*	-0.118	—
SHRSTc052	CA798030	F: TTTTAGAGCATTCCACITGCCCT R: CCATIGTTTCCACACTGAGAG	60	(TC) <sub>7</sub> N <sub>27</sub> (TC) <sub>3</sub> T <sub>15</sub> (TA) <sub>11</sub> (TTO) <sub>3</sub>	18	211–244	11	0.876	0.778	0.115	47a, 61

Table 1 *Continued*

Locus	GenBank Accession <sup>†</sup>	Primer sequence (5'-3')	T <sub>a</sub> (°C)	Repeat(s)	n	Size range (base pairs)	No. of alleles	H <sub>E</sub>	H <sub>O</sub>	f	LD‡
SHRSTc053	CA798178	F: TTCCCTTTCTTCTCTCTCTC R: AGTCGTTGCTACTGCTGG	56	(CT) <sub>18</sub> N <sub>5</sub> (CT) <sub>4</sub> TCT	22	198–224	7	0.703	0.364*	0.489	—
SHRSTc054	CA798214	F: CGATTGATGGTATTGGCTCTT R: TCACAGCTACGAATGGAA	60	(CT) <sub>2</sub> G(CTT) <sub>3</sub> CCT(CTT) <sub>7</sub>	22	82–100	3	0.592	0.727*	-0.235	—
SHRSTc055	CF972846	F: TCTTCTCTTTCCCCATTCCC R: CATCTCTTTCAAAACGCCA	60	CTC(TTC) <sub>3</sub> N <sub>10</sub> (TTC) <sub>8</sub>	22	209, 212	2	0.460	0.682*	-0.500	—
SHRSTc056	CF972885	F: ACCCTTTTGCCACCTTCTG R: CTTGACTTAAGTGTCCATTACACC	60	(AAG) <sub>7</sub>	22	95–107	4	0.504	0.500	0.009	48, 61, 64
SHRSTc057	CF972909	F: AGCGAAGCATATAATCATAGC R: GGCAATGATGGATACGACTAC	60	C.(TCA) <sub>8</sub> N <sub>13</sub> (TGC) <sub>2</sub> N <sub>23</sub> (TCA) <sub>4</sub>	22	143–149	3	0.487	0.182*	0.632	—
SHRSTc058	CF973870	F: GCTGTAGAGATTATTCTTTCGTC R: CCAAGAACAAAGAACCCA	60	(TCT) <sub>4</sub> (CTT) <sub>2</sub> CT(CAA) <sub>2</sub> N <sub>4</sub> (TTC) <sub>4</sub>	22	202–208	3	0.588	0.636*	-0.085	—
SHRSTc059	CF974239	F: ATGTGACGACCTCGATGA R: ACCAACCCCGAACAGT	60	(CTT) <sub>2</sub> T(GCC) <sub>2</sub> T <sub>4</sub> (TCA) <sub>9</sub>	22	93–105	3	0.563	0.591	-0.050	47a, 60, 61, 64
SHRSTc060	CA798467	F: TCGACTCGTTCGTCAAA R: CCCTTTATCCCTGGAGCA	60	(AAG) <sub>10</sub>	22	93–102	3	0.458	0.540	-0.197	45, 47a, 59, 70
SHRSTc061	CA796242	F: TCAACCGACCGACGAATAC R: AATCTCTACCCCGCTGGAG	60	(CAG) <sub>5</sub> CGGN <sub>9</sub> (CGG) <sub>2</sub> (CAG) <sub>5</sub>	22	179–217	4	0.580	0.591	-0.019	48, 50, 52, 56, 59, 64
SHRSTc062	CA795200	F: AGCCACAAAGCGTAGAG R: CAGCAAAGGGAGATCAGTC	60	(AAG) <sub>8</sub>	22	232–241	3	0.382	0.182*	0.529	—
SHRSTc063	CA797237	F: CTGTTCTTGCCCCCTGTT R: TGCTGTTCTCTTCTTG	60	(GT) <sub>3</sub> C(CTG) <sub>6</sub>	22	229–239	3	0.090	0.091	-0.012	45, 66
SHRSTc064	CF973954	F: TCCATACATTCTGCACCC R: TCGAGGAAAAGCTCTTACACT	56	(C) <sub>2</sub> AA(TA) <sub>3</sub> N <sub>5</sub> TATT(TA) <sub>9</sub>	22	291–387	11	0.642	0.591	0.081	48, 56, 59, 61, 68
SHRSTc065	CA796357	F: TCAAAGCAAACCCAGAGAAG R: TCGAAACAAAAACAAACCC	60	(GT) <sub>2</sub> (CT) <sub>2</sub> (AT) <sub>2</sub> AC(AT) <sub>11</sub> GT(AT) <sub>2</sub>	18	236–251	5	0.706	0.222*	0.692	—
SHRSTc066	CA798247	F: ACAGGAATCCCCATCAGCGA R: GCAATGACAGGCATGAGAGAG	60	CTTT(TC) <sub>15</sub>	22	210–224	7	0.732	0.727	0.007	50, 63, 68, 70
SHRSTc067	CF973890	F: GCTGGTGGAAAGATGGTAGAGA R: CCCGAAGAACCTAACGA	60	(TAAA) <sub>4</sub> N <sub>6</sub> T <sub>16</sub>	21	236–239	3	0.577	0.524	0.095	45, 47a, 50
SHRSTc068	CF972661	F: CTCCAACATCTCTACTCCATC R: GCTAAATCATAACAGCAACATCCACA	60	(AT) <sub>3</sub> (AAAT) <sub>4</sub> (AAGCA) <sub>2</sub>	22	330, 350	2	0.512	0.455	0.114	47a, 50, 64
SHRSTc069	CF974316	F: GGTGATTGAGATGAGAACAAAGGT R: CACAAGGGTAAAAAGAGAGAGAGA	60	(TC) <sub>2</sub> (CT) <sub>3</sub> G(T) <sub>3</sub> (C) <sub>3</sub> (CT) <sub>2</sub> G(TC) <sub>8</sub>	22	194–208	4	0.644	0.540*	0.156	—
SHRSTc070	CF974377	F: GATTACAACACCTTCTTCACCTAC R: ATCAGTTAACCCCTCC	56	(AG) <sub>6</sub> G(GA) <sub>5</sub>	22	223, 225	2	0.359	0.454	-0.273	47, 48, 50, 60
Mean							4.4	0.571	0.526	0.061	

T<sub>a</sub>, annealing temperature; n, number of accessions successfully genotyped out of 22; H<sub>E</sub>, expected heterozygosity; H<sub>O</sub>, observed heterozygosity; f, estimate of fixation index.

\*departs significantly from HWE at P < 0.05.

†GenBank Accession number given for an EST that contains both primer sites and the microsatellite region.

‡SHRSTc# given for loci showing significant linkage disequilibrium (LD) (P > 0.05), excluding loci that depart significantly from HWE.

**Table 2** Potential utility for mapping purposes

Parent 1	UF273	KA2-101	Sca6
Parent 2	Pound 7	K82	ICS1
Population type	F <sub>1</sub>	F <sub>1</sub>	F <sub>2</sub> *
No. of individuals	256	511	154
Segregating phenotype	Black pod and frosty pod diseases	Black pod and vascular streak dieback diseases	Witches' broom disease
No. of informative loci	19	17	19
Loci with expected 1:1 segregation ratio	16	12	—
Loci with expected 1:2:1 segregation ratio	0	5	19
Loci with expected 1:1:1:1 segregation ratio	3	0	—

\*created from selfing TSH516.

Dias LAS (2001) Chapter 12. The contributions of breeding. In: *Genetic Improvement of Cacao (Melhoramento Genético Do Cacaueiro)* (ed. Dias LAS), FUNAPE-UFG, Brazil. Translated into English by Abreu-Richart CE (<http://ecoport.org>).

Lewis PO, Zaykin D (2001) GDA (genetic data analysis): computer program for the analysis of allelic data, version 1.1. Free program distributed by the authors over the internet <http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php>.

Motamayor JC, Risterucci AM, Lopez PA *et al.* (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity*, **89**, 380–386.

Motamayor JC, Risterucci AM, Heath M, Lanaud C (2003) Cacao domestication II: progenitor germplasm of the *Trinitario cacao* cultivar. *Heredity*, **91**, 322–330.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population

genetics software for exact test and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Schnell RJ, Olano CT, Brown JS *et al.* (2005) Retrospective determination of the parental population of superior cacao (*Theobroma cacao L.*) seedlings and association of 2 microsatellite alleles with productivity. *Journal of the American Society for Horticultural Science*, **130**, 181–190.

Souza CAS, Dias LAS (2001) Chapter 1. Environment improvement and socio-economy. In: *Genetic Improvement of Cacao (Melhoramento Genético Do Cacaueiro)* (ed. Dias LAS), FUNAPE-UFG, Brazil. Translated into English by Abreu-Richart CE (<http://ecoport.org>).

Turnbull CJ, Butler DR, Cryer NC *et al.* (2004) Tackling mislabelling in cocoa germplasm collections. *INGENIC Newsletter*, **9**, 8–11.

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology*, **23**, 48–55.